

This is the author's final version of the contribution published as:

Mario Cataldi; Rossana Damiano; Vincenzo Lombardo; Antonio Pizzo.  
Lexical Mediation for Ontology-based Annotation of Multimedia. Springer.  
2012. pp: 113-134.

in

Oltramari, O.; Vossen, P.; Qin, L.; Hovy, E  
New Trends of Research in Ontologies and Lexical Resources

The publisher's version is available at:

<http://link.springer.com/content/pdf/10.1007/978-3-642-31782-8>

When citing, please refer to the published version.

Link to this full text:

<http://hdl.handle.net/2318/124954>

# Lexical Mediation for Ontology-based Annotation of Multimedia

Mario Cataldi, Rossana Damiano, Vincenzo Lombardo and Antonio Pizzo

**Abstract** In the last decade, the annotation of multimedia has evolved toward the use of ontologies, as a way to bridge the semantic gap between low level features of media objects and high level concepts. In many cases, the annotation terms refer to structured ontologies. Such ontologies, however, are often light scale domain oriented knowledge bases, whereas the employment of wide, commonsense ontologies would improve interoperability and knowledge sharing, with beneficial effects on search and navigation. In this chapter, we present an approach to the semantic annotation of media objects through a meaning negotiation approach that requires natural language lexical terms as interface and employs large scale commonsense ontologies. As a test case, we apply the annotation to narrative media objects, using a meta-ontology, called Drammar, to describe their structure. We present the annotation schema, the software architecture for integrating several large scale ontologies, and the lexical interface for negotiating the ontological term. We also describe an evaluation of the proposed approach, conducted through experiments with annotators.

---

Mario Cataldi  
Università di Torino, Italy e-mail: cataldi@di.unito.it

Rossana Damiano  
Università di Torino, Italy e-mail: rossana@di.unito.it

Vincenzo Lombardo  
Università di Torino, Italy e-mail: vincenzo@di.unito.it

Antonio Pizzo  
Università di Torino, Italy e-mail: antonio.pizzo@unito.it

## 1 Introduction

The huge amount of available multimedia resources requires novel forms of content indexing that is oriented toward re-use and retrieval. Beside the recent trend of user-generated annotations, structured semantic annotation has been proposed as a means to develop advanced search and retrieval tools, that rely upon both textual descriptions of the resource and signal content. In the last decade, thanks to the standards developed by the Semantic Web project, metadata can be expressed by reference to ontologies, thus guaranteeing the use of a shared, machine-readable format that goes beyond the limitations of keyword annotation. Ontologies are essential to represent and reason about shared meanings [22, 21] and allow the systems to describe the same resources with the same concepts belonging to a shared knowledge base [6].

Most approaches to the semantic annotation of multimedia content, aimed at bridging the so-called semantic gap by mapping low-level features onto semantic concepts, refer to specific sets of semantic descriptors, developed for specific content types and tasks. For example, consider the the MediaMill set of 101 semantic descriptors, suited for the MediaMill repository[47], or light ontologies such as LSCOM (a few thousands of concepts), specifically designed for a corpus of broadcast news[34]. Such approaches work for limited scale ontologies, where declarative rules and indexing algorithms directly refer to ontology nodes. On the contrary, when dealing with commonsense knowledge, the size and complexity of the ontologies make the mapping between low level features and ontology nodes hard. In order to support the use of large-scale commonsense ontologies in semantic annotation, we claim that the manual or semiautomatic generation of annotations is a crucial step: it provides training data for knowledge acquisition and learning [35] and ground truth data for evaluation purposes.

This paper presents a Wordnet-based lexical interface to the annotation, i.e., a system that permits a human user to access – via the lexical knowledge incorporated in WordNet – vast ontological knowledge bases for annotation purposes. Ontology concepts are selected by inserting natural language terms in a web-based system that helps the user visualize the multimedia documents and “negotiate” an ontological concept through a presentation of the glosses associated. This “meaning negotiation” process relies on the lexical knowledge-bases MultiWordNet [40] and WordNet[33]. The large-scale commonsense ontologies are the Suggested Upper Merged Ontology (SUMO, [38]) and Yet Another Great Ontology (YAGO, [50]), merged into YAGOSUMO project [32]. YAGOSUMO incorporates almost 80 millions of entities from YAGO (which is based on Wikipedia and WordNet) into SUMO, a highly axiomatized formal upper ontology. Thus, within the proposed framework, taking as input the word senses, the system queries YAGOSUMO in order to retrieve the ontological concepts that best match a set of ontological conditions imposed through YAGOSUMO properties. The description of situations, processes, and events require the connection of several concepts in a single relation. For this annotation, we rely upon the frame notion provided by the knowledge base FrameNet [2].

The lexicon-based approach described here is part of the CADMOS project, aiming at a Character-centred Annotation of Dramatic Media Objects (i.e., media objects having as their content character-enacted stories). We present the software architecture of the CADMOS project and the result of a test over an experimental corpus of narrative media (cf. [9]), where stories are presented in audiovisual and textual form. We believe that narrative media provide a valid test bed for the use of commonsense knowledge: notwithstanding the constraints posed by media and genres, they take as their object the real world, suitable to test the use of large commonsense ontologies.

The paper is organized as follows: in Section 2 we survey the relevant literature on the use of ontologies for multimedia content and the semantic annotation. Then, after the introduction of our case study, namely the annotation of dramatic media objects and the meta-ontology of narrative features, in Section 4.1 we introduce the architecture of the proposed framework and describe the methods and modules for implementing the lexicon-based method for the selection of the ontological concepts (Section 4.2). Finally, in Section 5 we report the experimental test, with user studies and analyses on the lexicon-based method for accessing the ontological knowledge base.

## 2 Related Work

In this section, we consider video annotation as a paradigmatic case for media annotation, both for the interest it has raised in the multimedia community [26] and for its relevance to the case study we describe in this paper.

Semantic annotation of video is generally performed by classifying content elements according to some ontology that represents its typical content [4]. Standardized metadata vocabularies, such as the LSCOM initiative [34], have been created to make the representation of video content interoperable, together with specialized vocabularies for videos related to various domains.

The annotation process implies a mapping of the individual elements of the video onto the terms of the reference ontology. The detection of the individual elements can be performed manually or automatically, through software systems for image and video analysis. The mapping of individual elements onto ontology concepts can be accomplished by simple pre-defined correspondences or through the definition of rules that establish relationships between the annotated terms to specify more abstract concepts. In this case, the terms of the ontology are mapped onto appropriate knowledge models that encode the spatio-temporal combination of low- or intermediate-level features [28, 16, 5]. The Video Event Representation Language (VERL) models events in the form of changes of state [18], following the paradigm of the event calculus [27]. This language introduces a compositional approach, yielding complex events from primitive concepts. It gives prominence to perceived objects and events, allowing for sequences or multi-threaded compositions, connected to the video through the beginning and end keyframes of the

event. The VERL approach does not refer to large-scale domain ontologies or to acknowledged patterns to provide a structure to the event models. Ballan et al. use the hierarchical linguistic relations over lexical entries encoded in WordNet to learn and refine rules that detect complex events from simple ones [3]. An ontology-based approach to the detection and annotation of events in video is pursued also by the Mind's Eye project [11]. In this work, the events detected in video are described as “verbs”, described in terms of a spatial model of motion. This approach relies on the paradigm of Ontological Realism, according to which the representation of the universals shared by different domain descriptions and applications is kept distinct from the representation of domain-specific templates. Beside the annotation of event, there is a growing interest for the representation of actions carried out by humans in a video (see, e.g., [52]). The representation of actions can be useful in the annotation of complex events, and can address many practical tasks, such as video surveillance.

An important limitation of current approaches is that they generally manage a limited range of concepts because of the inability to automatically recognize a wide range of elements from videos. In order to avoid these problems – and enable the use of a wider range of terms –, some annotation tools (as in [45]) allow the user to manually map a term with a specific ontological concept. The importance of the lexicon design for the task of recognition has been also pointed out by [23]: according to [23], the key to the creation of general-purpose content annotation and retrieval tools is the identification of a large lexicons and taxonomic classification schemes. The use of large-scale ontologies, however, introduces a new problem: the access to the data is, for the user, an extremely hard task, because of the size and the complexity of the considered data (cf. [34] and successive developments). An approach to improve the interoperability of the annotations is to constrain the scope of the semantic model: for example, the Lode ontology [29] describes the concept of public event (concert, performance, . . .), its structure, and properties, by abstracting on the descriptions provided by different directories.

A number of research projects directly address the problem of efficiently annotating video resources through large, shared, knowledge bases. Among all, the Advène project [43] addresses the annotation of digital video fragments by proposing a system that leverage free textual description of the content, cross-segment links, transcribed speech, etc. This information can be exploited to provide advanced visualization and navigation tools for the video. As a result of the annotation, the video becomes available in hypertext format. The annotation is therefore independent from the video data and is contained in a separate package that can be exchanged on the net.

A media independent project is provided by the OntoMedia ontology [24], exploited across different projects (such as the Contextus Project [25]) to annotate the narrative content of different media documents, ranging from written literature to comics and tv fiction. The OntoMedia ontology mainly focuses on the representation of events and the order in which they are exposed according to a time line, rather than to the specific features of the single medium (video, text, etc.). Rather than being tailored to event detection or annotation, OntoMedia lends itself to the

comparison of cross-media versions of the same story (for example, a novel and its filmic adaptation), where the story is rearranged according to different timelines in the different realizations of the story.

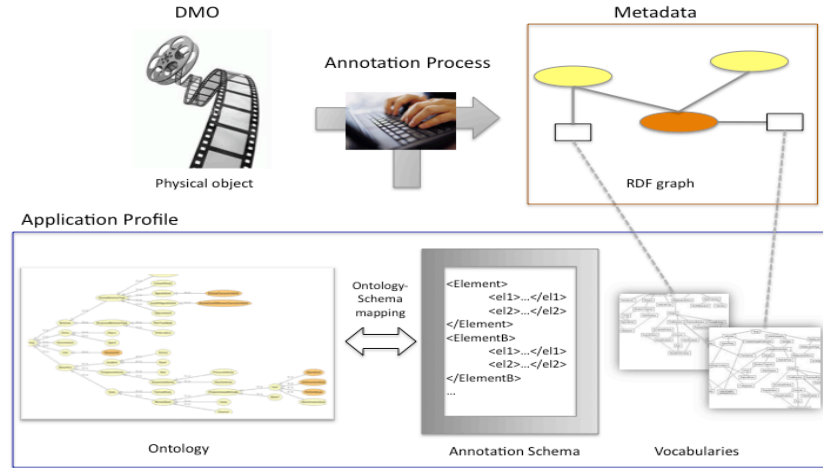
### 3 Case Study: Annotating Stories in Video

Narrative annotation requires the use of a semantic model to structure the description of stories. In order to make the annotated data interoperable and shareable among different projects, this models should abstract from the specific medium by which the story is conveyed and from the constraints posed by the conventions of specific genres or formats. In the Cadmos project, the annotation model is provided by the Drammar ontology.

Written in the Ontology Web Language [31], Drammar is not exclusively aimed at video, but relies on the concept of ‘dramatic media’, i.e., media displaying live action [17], that assign the character a primary role in the exposition of content. According to [44], in fact, media are more and more exploiting the power of narrative. With respect to the approaches presented in the previous section, Drammar shares with them the basic assumption that a media object can be segmented into meaningful units. However, it replaces the previous definition of units, respectively based on production (Answer project), thematic (Advène project) and structuralist concepts (OntoMedia ontology), with a segmentation methodology that relies on the identification characters’ actions.

In order to describe the behavior of characters, Drammar borrows the definition of agents from the BDI (Belief Desire Intention) model [42, 12], inspired by the framework of bounded rationality [8]. According to this model, agents devise plans (i.e., intentions) to achieve their desires, given their subjective beliefs about the current state of the world. This model, widely used in computational storytelling [36, 1], in Drammar is augmented with the notions of emotional states and moral values [39, 15, 14], to address the specific commitment of drama towards these notions.

Notice that the semantic model only describes the universe of discourse of drama. However, since the drama elements are also physical and abstract entities such as characters, institutions, objects, and so on, the annotation process needs a vocabulary for describing the real world counterparts of these elements. The paradigm of linked data [7] offers a way to link external semantic resources when describing some entity in an ontology. In the World Wide Web, classes, properties and individual of any ontology can be referred anywhere by using URIs to identify them. Thanks to this mechanism, in semantic annotation an external ontology can be employed as a terminological base without requiring an explicit integration of it in the annotation model. Cadmos relies on the paradigm of linked data to refer to individuals that belong to different datasets. For example, for describing the type of the objects that appear in a story, the Drammar ontology employs the *type* property. In each triple where this property is employed (<object,type,URI>), its value (the

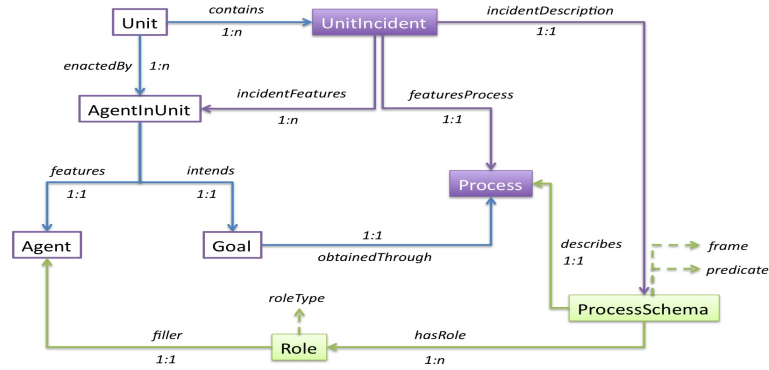


**Fig. 1** The annotation process framework. The annotation system incorporates the semantic model (Application Profile, below) and the external vocabularies, thus enforcing the correctness of the metadata encoded by hand by the human annotators and their translation into a formal language.

third element of the triple) is the URI of a concept in another ontology that corresponds to the type of that object. So, if the object is a car, the type property of this object will take as its value the URI of the concept of car in the external ontology that provides the vocabulary for the annotation of object types.

The schema depicted in Figure 1 represents the elements involved in a semantic annotation framework. The input to the process is given by the resource to be annotated (in Cadmos, a dramatic media object, or DMO) and the Application Profile. The Application profiles include the semantic model (in Cadmos, the Drammar ontology), the annotation schema and a set of vocabularies (i.e., external ontologies). The annotation schema is a hierarchical structure of descriptors, mapped onto the concepts represented in the semantic model; the values for the descriptors are given by the entries in the vocabularies.

The annotation process is accomplished by a human annotator with the help of a software tool that incorporates the Application Profile. Through this software, the annotator fills the annotation schema, selecting values for the descriptors from the vocabularies. Once the annotation schema has been filled, the system maps it onto the appropriate concepts and relations in the model, creating the right instances of the ontology classes and relations. The creation of the ontology instances is carried out by the system in a transparent way to the user: the output of this process is the metadata of the input DMO, encoded as an RDF graph. Also, in our framework, the selection of the values for the descriptors is not carried out by the annotator by direct access to the vocabularies (i.e., browsing the external ontologies) but is mediated by the natural language, as described in Section 4.



**Fig. 2** The template for annotating story incidents within the Cadmos system; the incident is described by an ontological concept, a semantic frame and its participants.

The top level of the Drammar ontology consists of five disjoint classes: *Unit*, *Dynamics*, *Entity*, *Relation* and *DescriptionTemplate*. The notions of unit, dynamics and entity generalize over a tripartite model of drama composed of plot tree (*Unit*), story advancement (*Dynamics*) and character (*Entity*), respectively [10]. According to this model, a story is segmented into units; units feature entities, involved in actions and events, i.e. the incidents that occur in units; units are arranged at different levels of detail, forming a tree structure. The *Dynamics* class contains the basic concepts for modeling the advancement of drama as a sequence of states interconnected by incidents. Finally, the *Relation* class subsumes the concepts that describe the properties of drama entities in a certain unit, such as the characters' goals and the conflicts among them. As stated before, agents are described according to paradigm of intelligent agents, following the Belief Desire Intention (BDI) model, as operationalized in several agent architectures [41, 37], and enriched with emotions and moral values [39, 13].

The annotation process centers on the description of the story units: a unit is enacted by certain characters, who perform actions in it, and/or contains certain naturally occurring events. As a result of these actions and events (collectively named incidents), the unit brings the world state from an initial state to a final state. In a situation calculus perspective [30], a unit can be seen as an operator characterized by preconditions and effects, that bridges the story world from a state in which the preconditions hold to one in which the effects hold. So, in Drammar, the unit is modeled as having preconditions and effects. The relation between the unit and the world state (before and after that unit) is modeled by the *hasPrecondition* and *hasEffect* properties, that connect the Unit with a StoryState.



As represented by the Figure 2, a Unit contains (*containsEvent*) some *UnitIncident* (an agents action or an event) and is *enactedBy* some Agents. The *UnitIncident* class (inspired by the Time Indexed Situation and the Time Indexed Participation patterns defined by [19, 20]) connects the occurrence of an event (no matter if it is an agent's action or a naturally occurring event) with the entities (agents and objects) which participate to it, and to set it into the time extent provided by a unit. Similarly to the *UnitIncident* class, the *StoryState* class connects the occurrence of a state (be it a mental state or a state of affairs) with the entities (agents and objects) which participate to the state, and to set this event in relation to a unit. The linguistic description of the incident, then, is attached to the *ProcessSchema* (or *StateSchema*) class, which in turn is connected to the entities which play a role in the incident through the *Role* class. The *Process* class is connected to the *ProcessSchema* class through the *incidentDescription* property (Fig. 2, below).

The *ProcessSchema* class describes the process through the following properties:

- The *predicate* data property links the *ProcessSchema* to a single concept represented into an external ontology of processes.
- The *frame* data property links the *ProcessSchema* to a single linguistic frame.
- The *hasRole* object property links the *ProcessSchema* to a thematic role (an instance of the *Role* class) belonging to the linguistic description of the process. Since a process normally encompasses multiple roles, an instance of *ProcessSchema* normally has multiple instances of the *hasRole* property.

The *Role* class represents a thematic role in the description of the process and can be filled by a drama entity through the *filler* property. The *roleType* property of the *Role* class provides a label for the type of role. By using this schema, the description of the process is entirely delegated to external ontological and linguistic resources, lifting Drammar from the responsibility of modeling common sense knowledge with which it is not concerned.

In order to better understand the final output of the RDF annotation, here we also provide a short example related to Act I, Scene 2, of Shakespeare's *Romeo and Juliet*. In particular, with our annotation system it is possible to describe the scene, where Romeo is entering, unseen, the garden of the Capulet's villa to find out where is Juliet. The movie fragment shows Romeo in the act of entering the garden and approaching the indoor balcony by the poolside in order to find Juliet's room.

```
:romeo rdf:type :Agent, owl:NamedIndividual;
      :age "18"^^xsd:int;
      :name "Romeo Montague"^^xsd:string;
      :gender "male"^^xsd:anyURI.

:goalOfRomeoInUnit1 rdf:type :Goal,
      :obtainedThrough :processRomeo;
      :hasStatus :goalRomeoStatus;

:processRomeo rdf:type :Process, owl:NamedIndividual;
```

```

:schemaRomeo rdf:type :ProcessSchema, owl:NamedIndividual
               :describes :processRomeo
               :predicate "finding"^^xsd:anyURI,
               :frame "Arriving"^^xsd:string;

:goalRomeoStatus rdf:type :GoalStatus,
                        owl:NamedIndividual;
:goalState "active"^^xsd:string.

```

## 4 Accessing Large Scale Commonsense Knowledge through a Lexical Interface

In this section, we describe the CADMOS system for the annotation of story and characters in video and the mechanism for accessing large ontologies from lexical knowledge it encompasses. In the CADMOS system, in fact, the NLP-mediated approach to large ontologies that we propose is employed to help human annotators identifying the appropriate concepts when describing what characters do in video, their motivations and the environment in which the action takes place.

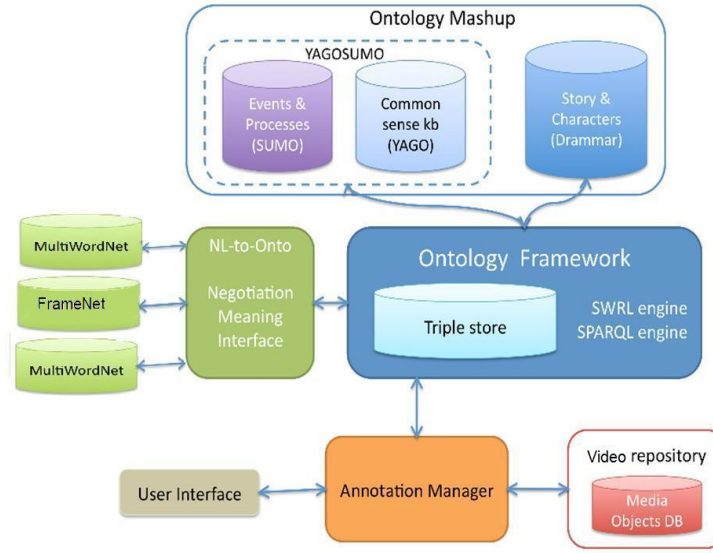
### 4.1 The Architecture of CADMOS

The architecture of CADMOS, illustrated in Figure 3, includes six main modules:

- the User Interface;
- the Annotation Manager;
- the Ontology Framework;
- the Ontology Mashup;
- the NL-to-Onto module;
- the Video Repository.

The system works as follows: the textual and multimedia documents to be annotated (also called media objects in this chapter) are stored and indexed within a repository, called Media Repository. In particular, video documents can be uploaded and visualized through a web-based User Interface, which is also the front-end for the annotation process. The Media Repository relies on a multimedia database to archive the video in the repository and a storage server to stream the requested video to the Annotation manager. The entire annotation work flow is led by the Annotation Manager which communicates with the Media Repository and the Ontology Framework, guiding the user within the annotation process.

The Ontology Framework carries out the reasoning services requested by the Annotation Manager and bridges the gap between the natural language input of the user and the ontological knowledge (Ontology Mashup). Currently, within the



**Fig. 3** The architecture of the Cadmos system.

proposed architecture, the Ontology Framework is provided by Jena<sup>1</sup> and it has been integrated with the Pellet reasoner<sup>2</sup>.

This mediation between natural language input and ontologies is possible through the use of the NL-to-Onto module: as explained in detail in Section 4, given a user input, expressed in one of the available languages, this module first permits to disambiguate the sense of the inserted term (in the selected language) by proposing to the user its different possible meanings; then, it associates in a transparent way the selected meaning to a unique sense in English. Moreover, when it is necessary, it permits to associate the selected sense to a semantic frame and to a set of thematic roles (therefore permitting a better contextualization of the annotated situation).

Currently, the Ontology Mashup contains two well-known ontologies: the Suggested Upper Merged Ontology (SUMO [38]<sup>3</sup>) and Yet Another Great Ontology (YAGO [50]<sup>4</sup>), merged into YAGOSUMO [32]<sup>5</sup>. This combined ontology provides a very detailed information about millions of entities, such as people, cities, organizations, and companies and can be positively used not only for annotation purposes, but also for automated knowledge processing and reasoning. The univocal mapping between a sense and an ontological concept is also possible thanks to the integra-

<sup>1</sup> <http://jena.sourceforge.net/>

<sup>2</sup> <http://clarkparsia.com/pellet/>

<sup>3</sup> <http://www.ontologyportal.org/>

<sup>4</sup> <http://www.mpi-inf.mpg.de/yago-naga/yago/>

<sup>5</sup> <http://www.mpi-inf.mpg.de/gdemelo/yagosumo.html>

tion of WordNet in YAGOSUMO<sup>6</sup>. The Ontology Mashup module also contains the annotation model (expressed by the Drammar ontology, described in the previous section), that provides the elements and properties employed to annotate the media objects within the system

It is important to note that the current architecture also support user queries on the annotated objects through the User Interface; in this case, the Ontology Framework translates the user request into a SPARQL query and performs the requested operation on the triple store (which contains the annotated information). The result is returned to the Annotation Manager that retrieves the relevant associated media objects and presents them to the user through the User Interface.

## 4.2 The Meaning Negotiation Process

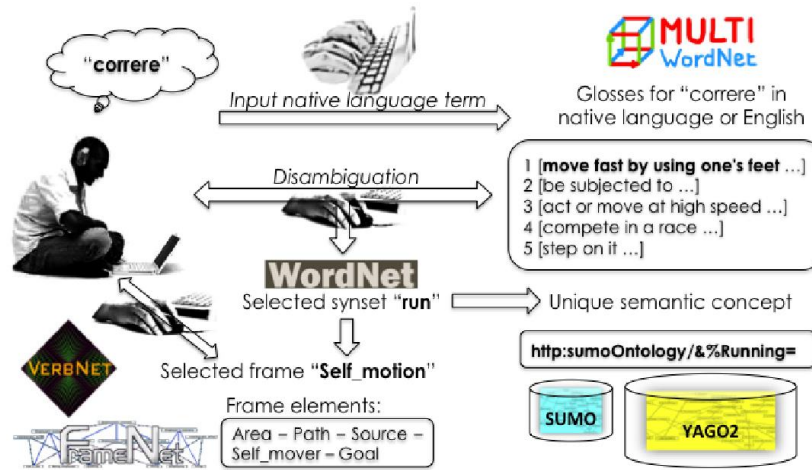
In order to fill the schema for describing story incidents with terms from external ontologies, our approach proposes a guided access to the ontology concepts based on natural language expressions. For this, we designed and implemented a tool that helps the user access the commonsense knowledge through a linguistic-based disambiguation process. The high-level schema of the entire work flow is shown in Figure 4).

In detail, the first part of this negotiation process can be described as a word sense disambiguation step aimed at associating to each natural language term/expression a unique definition which makes it distinguishable from any other possible meaning. In particular, for each element in the annotation schema, the system implements the following steps:

- the annotator initially expresses the content as a word (or a minimal set of words) in one of the available languages (English, Italian, Spanish, Portuguese, Hebrew and Romanian): the keyword-based query is forwarded to the NL-to-Onto module and the possible meanings of the query are shown by using the related glosses;
- the annotator disambiguates the meaning of her query by selecting the gloss that best matches her/his request;
- each gloss is then automatically and univocally mapped to a representative English WordNet synset;

---

<sup>6</sup> Each synset of WordNet becomes a class of YAGO [49]. They only exclude the proper nouns known to WordNet, which in fact would be individuals (Albert Einstein, e.g., is also known to WordNet, but excluded). Moreover, there are roughly 15,000 cases, in which an entity is contributed by both WordNet and Wikipedia (i.e. a WordNet synset contains a common noun that is the name of a Wikipedia page). In some of these cases, the Wikipedia page describes an individual that bears a common noun as its name. In the overwhelming majority of the cases, however, the Wikipedia page is simply about the common noun (e.g. the Wikipedia page Physicist is about physicists). To be on the safe side, they always give preference to WordNet and discard the Wikipedia individual in case of a conflict. This way, they can lose information about individuals that bear a common noun as name, but it ensures that all common nouns are classes and no entity is duplicated.



**Fig. 4** The disambiguation model proposed in this chapter includes several knowledge bases: MultiWordNet, WordNet, FrameNet and YAGOSUMO.

- finally, the system queries the YAGOSUMO knowledge base to retrieve the ontological concept that can positively represent the retrieved English synset.

More in detail, for each user query, the system retrieves the related definitions by querying the MultiWordNet data base searching for the synsets that are associated to the inserted query. In fact, within MultiWordNet, each synset (for each language) is represented as a tuple with four attributes:

- *id*: the WordNet synset identifier;
- *word*: the lemmas that can be associated to the considered WordNet synset;
- *phrase*: a locutionary expression that can represent the considered synset;
- *gloss*: a formal definition, as in a dictionary, expressed in natural language (with real examples), of the WordNet synset.

Thus, given the user's query, the system retrieves the related definitions by querying the NL-to-Onto module searching for the glosses which related "word" contains (also partially) the inserted term. This operation is initially performed on the table related to the user language. However, if related glosses are not available (in fact, except for the English table, it is not guaranteed a 1:1 mapping between each synset and a gloss), the system leverages the *ids* to retrieve, on the English table, the related English glosses (which are always guaranteed). At this step, the retrieved glosses are reported to the user (through the User Interface module) in her language (when available), or in English otherwise. The user then reads and analyzes the reported definitions in order to select the most suitable one.

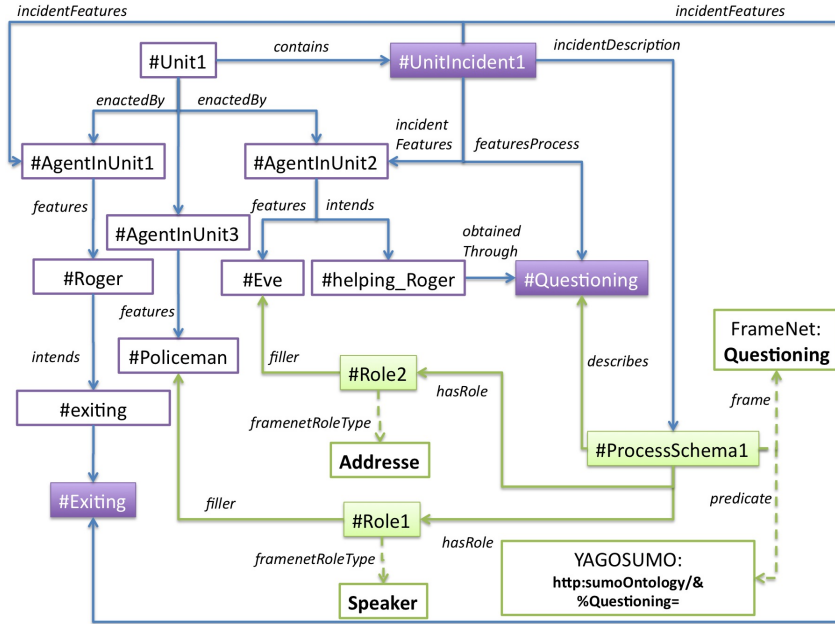
Then, the system leverages again the related synset *id* to retrieve additional information about the disambiguated sense. In particular, it is possible to query the

YAGOSUMO knowledge base to retrieve the related ontological concept; this is possible by using the ontological property *hasSynsetId*, represented within YAGOSUMO, which links an ontological concept to its related *id* in WordNet. In fact, the considered knowledge base has been constructed by merging, with an unsupervised method, the information expressed by Wikipedia (each article is represented as a class or an individual) and the linguistic hierarchical knowledge provided by WordNet. In fact, the information contained in Wikipedia is organized and structured based on its categorizations (that provides a basic hierarchical structure among the classes) refined and re-organized through the hyponyms/hypernym hierarchy provided by WordNet (i.e., they are converted into ontological high-level internal nodes). More in detail, YAGO has been automatically derived from Wikipedia and WordNet by including the taxonomic Is-A hierarchy as well as semantic relations between entities. The facts for YAGO have been extracted from the category system and the infoboxes of Wikipedia and have been combined with taxonomic relations from WordNet.

Note that YAGOSUMO, for each ontological concept, does not always associate the same *id* stored within MultiWordNet (this is because of data integration problems). Therefore, in order to avoid this problem, we leverage the related gloss to retrieve the correlated YAGOSUMO concept. This can be achieved through the ontological property *hasGloss* which links each single ontological concept to a unique formal definition extracted by WordNet. Note that, again, MultiWordNet and YAGOSUMO do not always associate the same gloss to each WordNet synset. In fact, they have been developed based on different WordNet versions and are therefore not completely aligned. Thus, when also this mapping system fails, our framework uses the related lemmas associated within MultiWordNet (stored in the “word” attribute) to retrieve the YAGOSUMO ontological concepts. This is possible through the use of the ontological property called “hasMeaning”, which links an ontological concept to all the terms (expressed as strings) that can be represented by the concept. Note that, using the associated lemmas, it could be possible to retrieve multiple concepts for each single selected definition. If this is the case, another negotiation step is required (i.e., the user needs to manually select the most suitable ontological concept).

Once the relevant concept has been retrieved, if the user is annotating a situation/event/action, the system can help the user in the annotation process by also proposing to the annotator the frame structure related to that concept (which can help describe the situation/event/action that needs to be annotated). Let consider for example the “Questioning” frame; it requires the specification of the elements “Message” (the exact wording of the questions), “Topic”, “Addressee” and “Speaker”. Using the information related to the frames, complex situations or events can be easily understood and annotated. The mapping between an ontological concept and a semantic frame is possible through the MapNet project [51] and FrameNet itself. When no frame is found (since the mapping is not yet complete), a generic frame is proposed to the annotator, accompanied by the set of 23 roles taken from the VerbNet project [46].

As an example of annotation, consider a scene from the classic Hollywood film “North by Northwest” by Alfred Hitchcock (1959). In this scene, the protagonist of the film, Roger Thornhill, gets off the train with Eve, disguised as a porter. The policeman who is pursuing Roger questions Eve about Roger and she answers that she has not seen him. Figure 5 illustrates the annotation of the incident. In the figure, *#UnitIncident1* features three agents (Roger, Eve and the Policeman, all instances of the *Agent* class). These agents are also the fillers (through the *filler* property) of the roles attached to the linguistic frame which describes the action featured in the incident (*#Questioning*). The role labels are provided by the *framenetRoleType* property, Speaker (Policeman, *#Role1*), Addressee (Eve, *#Role2*), Topic (Roger *#Role3*, not shown in the figure). The action is described by a SUMO concept (“Questioning”, see the *predicate* property) and by the ‘Questioning’ frame (*frame* property). The annotation also represents the characters’ goals (*#Exiting* for Roger, *#Helping\_Roger* for Eve). In Cadmos, the propositional content of goals, be it a state or a process, is also described through a situation schema, although this part has been omitted in the figure for space reasons.



**Fig. 5** The annotation of an example incident from North By Northwest (a policeman questions Eve about Roger).

## 5 Annotation test and discussion

In this section we report about an annotation test on a small-size corpus of narrative media objects. We analyze the numbers involved in our knowledge bases and the mappings between the lexical knowledge base and the ontological concepts and frames, respectively; then we report on the behavior of annotators through the “meaning negotiation” process and a comparison with free tagging.

### 5.1 Experimental Setting

The annotation of video through the ontological knowledge base challenges the sharing of the ontological concepts, by potentially introducing a large variety of terms both inter-language and inter-annotator, thus preventing interoperability. Thus, preliminarily, we measured the amount of positive mappings between the linguistic knowledge base terms (the terms stored in the NL-to-Onto module) and the ontology knowledge base YAGOSUMO. In particular, we tested how many terms contained within the lexical knowledge base (MultiWordnet) can be positively mapped to a concept within YAGOSUMO. In Tables 1(a) and (b) we report the results related to two different languages contained within MultiWordNet: English and Italian. As it is shown in Tables 1(a) and (b), 92.86% of the English terms reported in MultiWordNet (80.03% for Italian terms) are directly linked to an ontological concept. In fact, the presented system provides a guided access to ontological concepts related to  $\sim 95\%$  of the English verbs,  $\sim 86\%$  of the English nouns,  $\sim 90\%$  of the English adjectives and  $\sim 97\%$  of the English adverbs. The user can therefore leverage the expressiveness of the ontological knowledge base for a very significant percentage of natural language terms. Considering the Italian language, the percentage of terms that can be successfully linked to some ontological concept lowers a little (even if it remains higher than 75% for all the considered parts of speech); in fact, the considered ontologies (YAGOSUMO) are expressed in English and the system needs to find the correspondent concepts by also starting from glosses (or lemmas) in different languages. Thus, the data integration problems affect the mappings and lowers the percentage of terms in other languages associable to some ontological concept.

We also tested the mapping of MultiWordNet terms onto Framenet frames, that are employed for the annotation of situations/events/actions. Thus, we measured the percentage of natural language terms that can be positively mapped to a frame structure in FrameNet. As it is shown in Tables 2(a) and (b), nouns, adjectives and adverbs resulted in a very low percentage of positive mappings; as expected, verbs are more commonly considered for describing situations and events. In fact, for the verbs, our test reports a significantly higher percentage of positive mappings (60% for English and 70% for Italian). On the other hand, as explained in Section 4, when the system is not able to provide a mapping to a frame, it resorts to a general frame with high-level frame elements (taken from the knowledge base VerbNet).



	Total # Synsets	# Verbs	# Nouns	# Adjective	# Adverbs
Total # in MultiWordNet	102101	12144	68465	17917	3575
Total # of Mappings in YAGOSUMO	94817	10452	64831	16062	3472
Percentage	92.86	86.06	94.69	89.64	97.11

(a) English Terms

	Total # Synsets	# Verbs	# Nouns	# Adjective	# Adverbs
Total # in MultiWordNet	38653	4985	28517	3911	1240
Total # of Mappings in YAGOSUMO	30937	4332	21752	3643	1210
Percentage	80.03	86.90	76.27	93.14	97.58

(a) Italian Terms

**Table 1** Mappings among terms (English(a) and Italian (b)) in MultiWordnet and the considered large-scale knowledge base (YAGOSUMO).

	Total # Synsets	# Verbs	# Nouns	# Adjective	# Adverbs
Total # in MultiWordNet	102101	12144	68465	17917	3575
Total # of Mappings in FrameNet	22351	7193	10258	4352	548
Percentage	21.89	59.23	14.98	24.28	15.32

(a) English Terms

	Total # Synsets	# Verbs	# Nouns	# Adjective	# Adverbs
Total # in MultiWordNet	38653	4985	28517	3911	1240
Total # of Mappings in FrameNet	12357	3643	7252	1212	250
Percentage	31.96	73.07	25.43	30.98	20.16

(a) Italian Terms

**Table 2** Mappings among terms (English (a) and Italian (b)) in MultiWordnet and the the sematic frames stored in FrameNet.

The annotation experiment we ran asked to four users from different countries and speaking different languages, to annotate three different videos with the help of the annotation system. In particular, we considered the following videos:

- the 2-hour movie “North by northwest” (NbN), a classic Hollywood movie by Alfred Hitchcock, about an advertiser who escapes from both a criminal gang, who tries to kill him (having mistaken him for a CIA agent), and from the police, who tries to arrest him because of an unjust accuse of homicide;
- the multi-prized short animated movie “Oktapodi”, about an octopus who tries to save its partner from being cooked, after having been taken away from their love nest (a fish tank);
- a television commercial of the “Zippo” lighter, where a couple of gangsters try to burn a hostage but waste all the matches they have.

For all these resources, the users queried 289 times the lexical base for annotation. Considering all these requests, the users had to disambiguate in average among

2.83 glosses. It is interesting to note that this value is higher than the overall ambiguity; in fact, we calculated that, in average, each natural language term stored within our framework is associated to 1.71 glosses. This behaviour means that annotators tend to use terms that are more generic than the average (i.e., they results in a higher number of possible correlated definitions); in fact, more specific terms lower the average of this ambiguity factor to less than 2.

Moreover, we also asked the user to reply to a subjective qualitative-oriented questionnaire about the difficulty of using appropriate linguistic terms and the consequent selection of the adequate definition. For this, we asked the annotators to reply to the following questions:

- 1. Was it subjectively hard to make a selection from the list of definition provided by the system?
- 2. How many times did you revise your choice by searching for a synonym?
- 3. How many times did you change your interpretation because of the inadequate definitions proposed by the system?
- 4. How many times did you resort to free text, giving up the search on an ontological concept?

The users quantified the responses of the first question using a 3-point scale ratings (“easy to use”, “intermediate”, and “hard to use”), while for the other questions they simply counted the number of cases that were in accordance with the proposed questions.

## 5.2 Results and Discussion

Regarding the first question of the questionnaire, the users replied that the framework was “easy to use” in 80.23% of the cases, while for only 6.51% of the cases they found the system “hard to use” (for the 13.26% of the case they reported an “intermediate” difficulty), highlighting the simplicity of the proposed framework in annotating resources with ontological concept through a linguistic interface. Moreover, for 61.87% of the cases the user did not have to revise their query to search a suitable definition (second question), while in only 9.76% of the cases they had to repeat their requests (by inserting synonyms) more than once (and in 28.37% of the cases they reformulated their query only once).

It is important to note that, as already reported, some data integration problems emerged; in fact, regarding the third question, the user had to change their formulations in 38.76% of the cases, exhibiting the overall complexity of integrating different vast knowledge bases. In fact, in these cases, the annotators retrieved a set of results related to their queries but they were not satisfied with the proposed meanings; in other words, the system contained the terms provided by the users but they were not described in the way the users supposed. However, even with these problems, the users retrieved a satisfactory definition in 61.21% of the cases, exhibiting an overall robustness of the presented approach. Notice also that only in 16.92% of

Resource-based, 268 tags						
Title	Actor	Director	Production	Editing	Publishing	Genre
68	102	28	31	28	6	5

**Table 3** Resource-based tag analysis: number of tags per category.

the cases the users decided to resort to free text (fourth question) instead of insisting searching for the most suitable ontological concept. In other words, it means that in 83.08% of the cases the users easily retrieved a related ontological concepts in one or very few attempts.

About the behavior displayed with the annotation of the semantic relations for situations/events through frames (i.e., any video unit contains at least one event or action). In particular we asked to the user to report how many times they found a suitable frame. The answers to this question resulted in 62.45% of satisfactory mappings. It is interesting to note that the users typed in 97.67% of the cases a verb when they needed to annotate events, so a frame was likely present.

Finally, we checked whether our ontology-based annotation could be recovered, at least in part, from the free tags provided by users in the public repositories. So, we made an informal survey of the user-contributed tags on the feature film case (North by Northwest) in YouTube. After searching YouTube with the simple keywords “North by northwest”, we manually discarded all the results that did not belong to the original movie (59 % of the first 100 results consisted of advertising materials, CGI animations inspired by the movie, user-generated editings of the movie, etc.). We restricted our analysis to the Film & Animation category and considered only the first 100 results. By doing so, we collected 378 tags, yielding 183 different tags after eliminating the repeated tags. We then collected the tags of each result and manually analyzed them to let categories emerge, following the methodology of the Grounded Theory [48]. This methodology exploits both qualitative and quantitative aspects to group the data into categories and subcategories along the axis of each category, refining the categorization through the subsequent steps of analysis. Tags were divided into fourteen different categories, grouped into two main macro-categories: media-based tags, conveying information about media type, format, etc. and content-based tags. The latter can be further subdivided into actual content-based tags and general information about the resource, approximately corresponding to the Dublin Core data set<sup>7</sup> (information about the owner, the creator, the date, etc. ). The results of this analysis are illustrated in Tables 3 and 4. Most tags (268) belong to the description of the resource itself. Actual content based tags are only 29; 13 tags convey media information. Among the content based tags, most tags refer to characters (“Roger”, “mother”) or their qualities (“blonde”, “dress”). The “Other” category (49 unique tags) collects tags that are not related to the resource, such as advertising content, misspelled words, etc. .

<sup>7</sup> <http://dublincore.org/>

Content-based, 29 tags			
Character	Object	Environment	Action/Situation
10	8	7	4

**Table 4** Content-based tag analysis: number of tags per category.

Since a relevant number of tags are copied from the metadata that accompany the various editions of the movie, approximately one third of tags are proper names belonging to the production professionals (such as the director) and actors. Also, tags were multilingual, featuring, beyond English, German (6 tags) and Italian (1 tag). Finally, 26 tags were stop words, like the article “the” or the preposition “by”. Notice that this is due to the tagging interface of YouTube, that encourages users to slip multi-word tags (such as the title) into different tags. This informal analysis shows that, with respect to the story annotation schema we propose, the overlapping relies in the resource-based tags, that we encode according to the Dublin Core schema. The overlapping is not significant at the content level, that appears to be shallow in this example tagset. In particular, narrative aspects are mainly caught through the characters (10 occurrences) and the reference to objects (8 occurrences).

## 6 Conclusion

In this chapter we have presented an approach for the semantic annotation of media items, specifically targeted at video, that exploits very large scale, shared, common-sense ontology. The ontological terms are accessed through a linguistic interface that relies on multi-lingual dictionaries and action/event/situation template structures (semantic frames).

We have tested the validity and reliability of the proposed approach by allowing different users (not domain experts) annotate videos. The framework resulted promising from a user point of view because of its capacity to soften the complexity of accessing vast ontological knowledge bases. In fact, the presented application permits to leverage a large-scale commonsense knowledge base for annotating video by using semantic concepts. The access to such a component is provided by a multilingual linguistic interface, which revealed to be effective in the annotation task.

The future research plan includes an extension of alternative mapping systems among the different resources included within the proposed framework to help the user positively leverage a higher percentage of the natural language terms/expressions for annotation purposes. Moreover, we plan to extend the test of the proposed approach to a multi-lingual community of annotators to evaluate their feedbacks and collect wide-range annotations of different video sources.

## References

1. R. Aylett, M. Vala, P. Sequeira, and A. Paiva. Fearnot!—an emergent narrative approach to virtual dramas for anti-bullying education. *LNCS*, 4871:202, 2007.
2. C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.
3. L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE MultiMedia*, pages 80–88, October-December 2010.
4. M. Bertini, R. Cucchiara, A. Del Bimbo, and C. Torniai. Video annotation with pictorially enriched ontologies. In *Proc. of IEEE International Conference on Multimedia & Expo (ICME)*, Amsterdam, NL, July 2005.
5. M. Bertini, A. Del Bimbo, and G. Serra. Learning ontology rules for semantic video annotation. In *Proceedings of the 2nd ACM workshop on Multimedia semantics*, MS '08, pages 1–8, New York, NY, USA, 2008. ACM.
6. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 2009.
7. C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 4(2):1–22, 2009.
8. M. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge (MA), 1987.
9. M. Cataldi, R. Damiano, V. Lombardo, A. Pizzo, and D. Sergi. Integrating commonsense knowledge into the semantic annotation of narrative media objects. In *Proceedings of the 12th international conference on Artificial intelligence around man and beyond*, AI\*IA'11, pages 312–323, Berlin, Heidelberg, 2011. Springer-Verlag.
10. M. Cataldi, R. Damiano, V. Lombardo, A. Pizzo, and D. Sergi. Integrating commonsense knowledge into the semantic annotation of narrative media objects. *AI\* IA 2011: Artificial Intelligence Around Man and Beyond*, pages 312–323, 2011.
11. W. Ceusters, **J. J. Corso**, Y. Fu, M. Petropoulos, and V. Krovi. Introducing ontological realism for semi-supervised detection and annotation of operationally significant activity in surveillance videos. In *Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense and Security (STIDS)*, 2010.
12. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
13. R. Damiano and V. Lombardo. An Architecture for Directing Value-Driven Artificial Characters. *Agents for Games and Simulations II: Trends in Techniques, Concepts and Design*, pages 76–90, 2010.
14. R. Damiano and V. Lombardo. An Architecture for Directing Value-Driven Artificial Characters. *Agents for Games and Simulations II: Trends in Techniques, Concepts and Design*, pages 76–90, 2011.
15. R. Damiano and A. Pizzo. Emotions in drama characters and virtual agents. In *AAAI Spring Symposium on Emotion, Personality, and Social Behavior*, 2008.
16. A. Ekin and A. M. Tekalp. Automatic soccer video analysis and summarization. In *Storage and Retrieval for Media Databases*, pages 339–350, 2003.
17. M. Esslin. *The Field of Drama*. Methuen, London, 1988 (1987).
18. A. R. François, R. Nevatia, J. Hobbs, and R. C. Bolles. VerI: An ontology framework for representing and annotating video events. *IEEE MultiMedia*, 5:76–86, 2005.
19. A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *Proc. EKAW 2002*, Siguenza (SP), 2002.
20. A. Gangemi and V. Presutti. Ontology design patterns. *Handbook on Ontologies*, pages 221–243, 2009.
21. T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43:907–928, December 1995.

22. N. Guarino and P. Giaretta. Ontologies and knowledge bases: Towards a terminological clarification. In N. J. I. Mars, editor, *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32. IOS Press, Amsterdam, 1995.
23. A. Hauptmann. Towards a large scale concept ontology for broadcast video. *Image and Video Retrieval*, pages 2047–2047, 2004.
24. M. Jewell, K. Lawrence, M. Tuffield, A. Prugel-Bennett, D. Millard, M. Nixon, and N. Shadbolt. OntoMedia: An ontology for the representation of heterogeneous media. In *In Proceeding of SIGIR workshop on Multimedia Information Retrieval*. ACM SIGIR, 2005.
25. M. O. Jewell, K. F. Lawrence, M. M. Tuffield, A. Prugel-Bennett, D. E. Millard, M. S. Nixon, m. c. schraefel, and N. R. Shadbolt. Ontomedia: An ontology for the representation of heterogeneous media. In *Multimedia Information Retrieval Workshop (MMIR 2005) SIGIR*. ACM SIGIR, ACM SIGIR, 2005.
26. I. Kompatsiaris, S. Marchand-Maillet, R. van Zwol, and S. Marcel. Introduction to the special issue on image and video retrieval: theory and applications. *Multimedia Tools and Applications*, pages 1–6, 2011.
27. R. Kowalski and M. Sergot. A logic-based calculus of events. *New generation computing*, 4(1):67–95, 1986.
28. R. Leonardi and P. Migliorati. Semantic indexing of multimedia documents. *IEEE MultiMedia*, 9:44–51, April 2002.
29. X. Liu, R. Troncy, and B. Huet. Finding media illustrating events. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–8. ACM, 2011.
30. A. McCarthy. Mental situation calculus. *TARK: Theoretical Aspects of Reasoning about Knowledge*, 1986.
31. D. McGuinness, F. Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10:2004–03, 2004.
32. G. d. Melo, F. Suchanek, and A. Pease. Integrating yago into the suggested upper merged ontology. In *Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence - Volume 01*, pages 190–193, Washington, DC, USA, 2008. IEEE Computer Society.
33. G. A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41, 1995.
34. M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13:86–91, July 2006.
35. L. Nixon, S. Dasiopoulou, J.-P. Evain, E. Hyvonen, I. Kompatsiaris, and R. Troncy. Multimedia, broadcasting, and eulture. In J. Domingue, D. Fensel, and J. A. Hendler, editors, *Handbook of Semantic Web Technologies*, pages 911–975. Springer-Verlag, Berlin Heidelberg, 2011.
36. E. Norling and L. Sonenberg. Creating Interactive Characters with BDI Agents. In *Proceedings of the Australian Workshop on Interactive Entertainment IE2004*, 2004.
37. T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, editors. *Intelligent Virtual Agents, 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005, Proceedings*, volume 3661 of *Lecture Notes in Computer Science*. Springer, 2005.
38. A. Pease, I. Niles, and J. Li. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, volume 28. Edmonton, Canada, 2002.
39. F. Peinado, M. Cavazza, and D. Pizzi. Revisiting Character-based Affective Storytelling under a Narrative BDI Framework. In *Proc. of ICIDIS08*, Erfurt, Germany, 2008.
40. E. Pianta, L. Bentivogli, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January 2002.
41. A. Pokahr, L. Braubach, and W. Lamersdorf. Jadex: a BDI Reasoning Engine. *Multiagent Systems, Artificial Societies and Simulated Organizations*, 15:149, 2005.
42. A. Rao and M. Georgeff. Deliberation and intentions. In *Proc. of 7th Conference on Uncertainty in Artificial Intelligence*, Los Angeles, 1991.

43. B. Richard, Y. Prié, and S. Calabretto. Towards a unified model for audiovisual active reading. In *Tenth IEEE International Symposium on Multimedia.*, pages 673–678, Dec. 2008.
44. M. Ryan. *Avatars of Story*. University of Minnesota Press, 2006.
45. C. Saathoff, S. Schenk, and A. Scherp. Kat: The k-space annotation tool. In *SAMT 2008, Demo Session Proceedings*, 2008.
46. K. K. Schuler. *Verbnet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA, 2005. AAI3179808.
47. C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of ACM Multimedia.*, pages 421–430, Santa Barbara, USA, October 2006.
48. A. Strauss and J. Corbin. *Basics of qualitative research: grounded theory procedures and techniques*. Sage Publications, Newbury Park, Calif., 1990.
49. F. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM, 2007.
50. F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *WWW '07: Proceedings of the 16th International World Wide Web Conference, Banff, Canada*, pages 697–706, 2007.
51. S. Tonelli and D. Pighin. New features for framenet - wordnet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, CO, USA, 2009.
52. G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong. Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor. In *Proceedings of the ACM Multimedia Conference*, pages 165–174, 2009.